ABSTRACT

The most striking feature of Earth is the existence of life, and the most striking feature of life is its diversity, popularly known as – biological diversity or biodiversity. Biodiversity is the biological capital of our planet and it forms the foundation upon which the human civilization is built. It is fundamental for the Earth’s life support system. The history of human civilization and material culture and development of economic systems are all indirectly associated with the use and management of biotic and abiotic resources, together called as “natural resources”. These natural resources are often taken for granted which provide much essential natural service. Hence, it is in the interest of mankind that these resources are used in sustainable manner, cautiously, so as to ensure continued survival of human race on this planet. Efficient access to knowledge base on these natural resources and process is essential for their effective conservation and sustainable use.

The term "Biodiversity Informatics" was coined to circumscribe the application of information technology tools and technology to biodiversity information, principally at the organismic level. It thus deals with information capture, storage, provision, retrieval, and analysis, focused on individual organisms, populations, and taxa, and their interaction. It covers the information generated by the fields of systematics (including molecular systematics), evolutionary biology, population biology, behavioural sciences, and synecological fields ranging from pollination biology to parasitism and phytosociology. Biodiversity Informatics is considered a part of biological informatics sandwiched between - and strongly overlapping with - environmental informatics and molecular bioinformatics. It will provide the skeleton for a generalized scientific information infrastructure in biology. However, there is disparity and uneven distribution of biodiversity and biodiversity information across the globe. Similarly, the progress of biodiversity informatics is currently concentrated outside mega-biodiversity regions of the world. Thus, for mega-biodiversity developing nation such as India, it is essential that we realize the biodiversity informatics as cornerstone of our economic, ecological and social well being. Therefore, the aim of this exercise is to design and implement biodiversity information management system for mega-biodiversity nation such as India, with cataloguing Indian fauna as a case study.

Chapter 1: Biodiversity Informatics: A Review

Biodiversity informatics is emerging discipline, which deals with collection, collation, analysis, prediction, and dissemination of data and information related to biotic resources of the earth. This chapter reviews the global progress in the field of biodiversity informatics by grouping it in four categories, viz. (i) mobilizing biodiversity data, (ii) standards, protocols, and tools, (iii) informatics infrastructure building initiatives, and (iv) capacity building, outreach, and open access initiatives.

There are estimated to be 1500 biodiversity information resources. Since information about these resources is distributed, it is difficult to review the progress made in biodiversity informatics. Thus, BIR, Biodiversity Information Resources Database was developed. Analysis of BIR confirms that similar to uneven distribution of biodiversity, biodiversity information and informatics activities are unevenly distributed. While, biodiversity is concentrated within tropical mega-biodiversity regions which is both developing and under-developed; biodiversity information and informatics activities are concentrated in non mega biodiversity, developed world institutions. Majority of the biodiversity information resources are coarse in nature (global, regional, or national), and in English. There is a need for micro-focus databases in vernacular languages. Further, there needs to be an impetus on development of biodiversity information handling protocols, tools, and standards so to achieve interoperability amongst the resources. BIR analysis clearly emphasizes the need of biodiversity informatics activities in these regions, if our goal of sustainable use and conservation of biotic resources is to be achieved.

Biodiversity informatics activities are critically significant for the mega-biodiversity, developing economy such as India. Three strong points of argument for undertaking biodiversity informatics activities in India includes, (a) exploding population – a national challenge, (b) natural resources and economy, and (c) emerging biodiversity knowledge catastrophe. Form the review it appears that biodiversity informatics activities in India are in its nascent stage, and needs support.
Chapter 2: IndFauna, Electronic Catalogue of Known Indian Fauna

India is known to harbor 89,451 faunal species which is 7% of the world’s known faunal diversity. While, this estimate is nearly 9 years old, numerous new species have been described in recent past. However, there is no single repository where baseline data and information regarding these species could be accessed. Currently, this data is distributed with several individuals and institutions, and majority of the times it is in non-interoperable forms and format. In order to address this, I have conceived and developed the IndFuana, electronic catalogue of known Indian fauna.

While briefly reviewing the faunal diversity in India, I have attempted to reason the development of IndFuana. IndFauna collates following baseline information viz., (a) valid scientific name with authority, year of publication according to accepted taxonomic opinion, (b) systematics of the species from kingdom to forma level, (c) synonyms with authority and publication year, (d) common, local, or vernacular names with languages, and regions, (e) occurrence along with details of source data, (f) bibliographic details, and (e) multimedia artwork, etc.

Oracle 9i has been implemented to develop the database structure which is described in detail in this chapter. One of the unique features of IndFauna, is its entirely transparent process for which web based set of data management, and data curation modules. Another feature of IndFauna is LinkOut with other datasets such as sequence databases. LinkOut has also been provided with GoogleImages. Similar LinkOut could be established with other databases such as Google Scholar, Google Books, Barcode of Life Database (BOLD), Species2000 ITIS Catalogue of Life, uBIO, PubMed, Scirus, and ZooBank, etc.

Another unique feature of IndFauna is its data cleaning process, which is based on approach of both prevention and correction. Over 100+ self volunteered “Taxon Experts” contributed towards enhancing data quality, taxonomic authenticity and validity of collated data. Process of data cleaning and taxonomic scrutiny has been described in detail.

IndFauna currently collates baseline data about 94500 known Indian faunal species. It not only facilitates easy access to Indian faunal diversity data, but also would provide sound base for resolving conflicts in taxonomies, planning future
research and analysis. IndFauna, thus have strong potential for act as “central registry of names of organisms”, which could form the backbone of national, and regional biodiversity information system collating and disseminating data on host of other parameters and factors that are responsible for dynamics, and health of our natural resources and their habitats.

Chapter 3: IndFauna: Data Cleaning, Taxonomic Scrutiny, and Lessons Learnt

Key purpose of electronic catalogue of known organisms such as IndFauna is to collate and disseminate species occurrence data. The uses of species and occurrence data are wide and varied and encompass virtually every aspect of human endeavor – food, shelter and recreation; art and history, society, science and politics. However, efficient and effective applicability and use of this data depends on quality of data. Thus, data cleaning enhances the “fitness for use” of data. Since, IndFauna has collated data from secondary sources; an approach of “prevention and correction” was adopted for enhancing the quality of data. These approaches and development of data curation modules has been described in this chapter. Taxonomic scrutiny which forms part of correction approach of data cleaning was carried out using offline and online data curation modules.

IndFauna has also been subjected to quantitative and qualitative analysis. While quantitative analysis was aimed as identifying gaps in collated and accessible data, qualitative analysis was attempted to identify taxonomic discrepancies, and help resolve them. During IndFauna development, it was learnt that one of the major reason for these discrepancies is disparity in availability of nomenclature change literature to the taxonomists of the developing world and availability of taxonomic papers published by developing world scientists to their counterparts in developed part of the globe. However, development of electronic catalogues of names of known organisms would help in pointing out these issues. I have attempted to highlight a few of such discrepancies found while developing IndFauna, an electronic catalogue of known Indian fauna and comparing it with existing global and regional databases. These discrepancies can be grouped into three categories, viz., (a) hierarchical differences, (b) spelling differences, and (c) homonymies.

Resolving these discrepancies is a matter of taxonomic discussion. They need to be resolved using nomenclatural rules. However, through the examples quoted in the chapter, I am attempting to demonstrate the role of electronic catalogues in bringing issues or discrepancies to the knowledge of taxonomic community, starting a
dialogue between taxonomists across the globe and identifying issues of common concern. In order to notice such discrepancies and resolve them quickly, it is essential that a wrapper be developed which traverses through various electronic catalogues searching for taxonomic anomalies. This calls for increasing collaboration among the various electronic catalogues of names of known organisms, which is far from happening in developing world. Thus, it justifies the investment in development of national electronic catalogues as core of the biodiversity informatics activities in mega-biodiversity regions of the world.

Chapter 4: JaivaNaksha: Web mapping of Occurrence data and its Geo-referencing

One of the important goals of biodiversity databases is to provide location information of species in order to empower decision-making in context of planning, developmental projects, conservation and invasive species management. Most biodiversity data usually has been collected from older collections when accurate maps and global positioning systems were unavailable. Consequently, species in museum collection or older literature seldom are associated with geographic coordinates. Locality references in most cases are in form of textual descriptions. In this chapter, while detailing the development of JaivaNaksha, web mapping application for indFauna; I review the challenges of assigning geo-coordinates to such descriptive locations, and standards to be adopted while doing so, as majority of the locality records documented in current exercise falls in this category.

Thus, one of the important challenges while designing JaivaNaksha is to accommodate the various types of locality records. These ranged from general references like ‘Throughout India’ to ‘Maharashtra’ (state names) to ‘Sindhudurg district’ (district name) to point locations (villages, towns). Apart from such textual descriptions, there also exist references to both arbitrary and precisely defined regions, examples of which include ‘Southern India’ for the former and ‘Thar Desert’ or ‘Rajaji National Park’ for the latter. The third type of location data that was encountered was in the form of river/water body names including lakes, mangroves, lagoons and estuaries. Our philosophy has been to provide an accurate representation of the described localities without taking decisions on their appropriateness. Hence, although ‘Throughout India’ would be discounted today for most species, we have sought to plot the same on the map.
JaivaNaksha, a user friendly web-based geographic information system (GIS) has been developed considering all these complexities and based on all these issues, JaivaNaksha and its backend database schema was designed using a combination of open source and proprietary technologies. A primary challenge in its development was to dynamically create maps on the fly, for which PHP Mapscript and session variables were used. It further generate the report which provides detailed information for the species with respect to occurrence data, location type and the source in which that location has been described. The maps are expected to be further refined as more occurrence data is georeferenced. The process of georeferencing of IndFauna occurrence records is also described together with the data curation module developed for this purpose.

During development of JaivaNaksha, it was realized that there is resistance to share both spatial data as well spatial data products such as shape files. While the need for open access to such data and products has been emphasized several times in the past, there is no open access repository where such data and products could be contributed. Thus, Open Access Geospatial Data Repository (OAGDR) was developed to exchange / share shapefiles of commonly used geographic features.

**Chapter 5: National Biodiversity Information Infrastructure: Challenges, Potentials and Roadmap**

During development of IndFauna, and allied products described in earlier sections of the thesis, urgent need was felt to evolve planned mechanism to collect, collate, and disseminate data and information about Indian biodiversity. As noted earlier such data and information is currently distributed, isolated, in heterogeneous forms and format, and most seriously locked up in institutional and individual cupboards under the misconceptions of national security, intellectual property related sensitivity. While, there are few sporadic, isolated efforts being made in recent past, there is a need to coordinate these activities under one single over-arching umbrella. Thus, there is a need to conceive and establish “**National Biodiversity Information Infrastructure (NBII)**”. Current technological and political scenario presents ample scope to undertake establishment of such a facility that is capable of collation, analysis, and dissemination of biodiversity and ecosystem related information form / to distributed sources.

NBII should be an interoperable network of biodiversity databases, information networks and systems, traditional knowledge, peoples biodiversity
registers, and information technology tools that will enable users to navigate and put to use the nation’s vast quantities of biodiversity and ecosystem information to produce national economic, environmental, and social benefits. Thus, it would be an overarching information facility, which would leverage on progress made so far by the various information systems, networks and databases spearheaded by various individuals, institutions and groups within and outside India.

In this chapter, I have attempted to elaborate on the vision and operational objectives, potential work programs, major milestones, as well performance indicators of such a system. While highlighting the challenges, I have further discussed some of the technical implementation related issues such as use of web services architecture for evolving such an infrastructure, with its merits and demerits, as learned from similar implementations in other regions of the globe. Further, I have discussed the governance structure of NBII.

It is my belief that such a facility will contribute towards economic growth, ecological sustainability, and social outcomes through increasing the utility, availability and completeness of new and existing biodiversity and ecosystem information resources.

Annexure I: BIR, Biodiversity Information Resources Database

Our progress in biodiversity informatics is similar to that of the uneven distribution of biodiversity and biodiversity data. While, most of the development of information bases, standards and tools is happening in developed part of our globe, mega-biodiversity developing nations are lagging behind in collation and dissemination of data about their biodiversity. Currently, metadata of the biodiversity information resources themselves is not accessible at a single click of a mouse. To overcome this development of BIR, Biodiversity Information Resources database was undertaken. Annexure I, describe the development of BIR which has collated metadata for over 1300+ biodiversity information resources. Thus, BIR facilitate up-to-date and current documentation of existing and new biodiversity and ecosystem information resources. It was felt that metadata repository such as BIR needs to be constantly updated, if our goal is to bridge the imbalance between the biodiversity and ecosystem informatics products and distribution of biodiversity and its data.

Annexure II: Open Access Geospatial Data Repository (OAGDR)

In recent times, the need of having an easily accessible spatial data infrastructure has been emphasized by several communities such as scientists,
technologists, academicians, planners and even commoners. However, simple geospatial products such as shapefiles are not available when they are needed the most. To overcome this impediment, and foster a community-driven effort towards building a geospatial data infrastructure, an Open Access Geospatial Data Repository (OAGDR) has been developed. This annexure describe the purpose and development of OAGDR in detail.

**Annexure III: Connecting Diversity: Pilot project for development of an interoperable framework for connecting distributed and heterogeneous bioresources databases**

Chapter 5 of this dissertation deals with development of National Biodiversity Information Infrastructure (NBII). However, during my 15+ years of work in the area of biodiversity informatics, I was always queried if such a grand vision can ever be implemented? About two years back with support from Government of India’s Department of Biotechnology, I was awarded a pilot project to implement web services architecture to interconnect biodiversity and bioresources databases. Annexure III, deals with progress of this pilot project which has been able to harvest together over 500,000 records from 7 distributed and heterogeneous databases through 2 data providers. This data is accessible through IBIF prototype portal ([http://www.ibif.net.in/](http://www.ibif.net.in/)). Experience of this pilot project once again reaffirm that technology is not an impediment in bridging and interconnecting the data, but its mindset of individuals who hold this data.